On Prior Distributions For Binary Trials

by Seymour Geisser*

University of Minnesota

Technical Report No. 410

November 1982

*Supported in part by NIH Grant GM25271.

1. Introduction

It was pointed out by G. Barnard et al. (1962) that if it were reported that a coin whose probability of heads is θ came up heads t times and tails $n-t$ times in a series of independent tosses that, irrespective of the stopping rule, the likelihood was

$$L(\theta) \propto \theta^t(1-\theta)^{n-t} , \tag{1}$$

and the likelihood principle would then dictate that any inference about θ should not depend on which stopping rule was actually used. Two common stopping rules are: (a) Fix the total number of tosses and observe the number of heads (binomial sampling); (b) Observe the total number of tosses required to attain a fixed number of heads (negative binomial sampling).

In case (a) we have the sampling distribution of T the number of heads

$$\Pr[T=t|n] = \binom{n}{t} \theta^t(1-\theta)^{n-t} . \tag{2}$$

In case (b) the sampling distribution of N the number of tosses required to obtain t heads,

$$\Pr(N=n|t) = \binom{n-1}{t-1} \theta^t(1-\theta)^{n-t} \qquad n=t,t,t+1, \ \ldots \ . \tag{3}$$

Now there are Bayesians who have developed rules for obtaining reference prior distributions which purport to express little or no information regarding the parameter θ. Jeffreys (1961) invokes invariance; Box and Tiao (1973) recommend priors such that likelihoods are data translated in some sense; Akaike (1978) and Geisser (1979) formulate procedures which involve the predictive distribution and Kullback-Leibler divergence measures whilst Bernardo (1979) uses the notion of maximizing entropy in the limit, and Zellner (1977) maximizes the Shannon information in the data relative to that of the prior. All of the above

methods with the exception of Geisser's and Zellner's yield the same reference prior

$$p_B(\theta) \propto \theta^{-\frac{1}{2}} (1-\theta)^{-\frac{1}{2}} \tag{4}$$

for the binomial and

$$p_N(\theta) \propto \theta^{-1} (1-\theta)^{-\frac{1}{2}} \tag{5}$$

for the negative binomial case. Hence the posterior densities for these two cases are

$$\begin{aligned} p_B(\theta|t,n) &\propto \theta^{t-\frac{1}{2}} (1-\theta)^{n-t-\frac{1}{2}} \\ p_N(\theta|t,n) &\propto \theta^{t-1} (1-\theta)^{n-t-\frac{1}{2}} \end{aligned} \tag{6}$$

respectively. Zellner's method leads to priors that differ from the above and differ between themselves. Geisser's method does not provide unique solutions in these cases. In fact for all these methods the prior distribution will depend on the sampling rule and consequently so will the posterior distribution (although to a far lesser degree). Dependence on the sampling rule is defended by some Bayesians as can be adduced by the following quotes.

Box and Tiao (1973, p. 46): "In general we feel that it is sensible to choose a noninformative prior which expresses ignorance relative to information which can be supplied by a particular experiment. If the experiment is changed, then the expression of relative ignorance can be expected to change correspondingly."

Akaike (1980, p. 147): "The expected behavior of the likelihood function is certainly different for the two schemes . . . and it is <u>irrational</u> (my emphasis) to adopt one and the same prior distribution, irrespectively of the expected difference of the statistical behavior of the likelihood functions."

Bernardo (1979, p. 144): "Indeed, it is known that even from a purely personalistic point of view, one must integrate over the sample space to design an experiment. It does not seem unnatural to me that one has to do the same to analyze the implications of its results."

Zellner (1977, p. 231) "Since the purpose of a MDIP [maximal data information piror] is to allow the information provided by an experiment to be featured [in the posterior distribution], it seems natural that this form of a MDIP *pdf* that accomplishes this objective be dependent on the design of an experiment".

Jeffreys, who suggested the use of the root of Fisher's expected information as the prior density which several of the previously suggested methods reduce to in this case, interestingly enough appears to prefer a uniform prior on θ. He states (1961, p. 125) "Then there is no objection to the uniform distribution and no other . . . has been seriously suggested though there is something to be said for . . . [use of the square root of Fisher's expected information].

Although other Bayesians believe that the prior should not depend on the sampling distribution, which reflects the experiment, it is not inconceivable that the experiment itself induce the paradigm involving the parameter. More importantly the parameter is very often an unobservable construct mainly devised to foster a model which facilitates the prediction of future observations, so that a prior for the parameter may be a matter of convenience.

On the other hand, these general arguments for sampling dependent priors seem rather weak for this particular problem. We shall attempt in this paper to present an argument for a particular non-informative prior that adequately addresses the prediction problem and induces

a single prior for the various types of sampling as a consequence of a limiting argument. It turns out to be the uniform prior preferred by Bayes, Laplace and Jeffries when there is presumed to be no prior information.

2. The Uniform Prior — A Justification.

Of course the posterior density for a uniform prior density $p(\theta) = 1$, is

$$p_U(\theta|t,n) \propto \theta^t(1-\theta)^{n-t} \quad . \tag{7}$$

If we now addend to this the further problem of trying to predict the total number of heads R including the t heads already sampled, in a total of M trials (n already sampled), we calculate the probabilities that R = r given M and the alternative priors for θ:

$$Pr_B[R=r|M,n,t] = \frac{\binom{r}{t-\frac{1}{2}}\binom{M-r}{n-t-\frac{1}{2}}}{\binom{M}{n}} \qquad r=t,\ t+1,\ldots,\ \min(n,\ M-n+t) \tag{8}$$

$$Pr_N[R=r|M,n,t] = \frac{\binom{r}{t-1}\binom{M-r}{n-t-\frac{1}{2}}}{\binom{M-\frac{1}{2}}{n-\frac{1}{2}}} \tag{9}$$

$$Pr_U[R=r|M,n,t] = \frac{\binom{r}{t}\binom{M-r}{n-t}}{\binom{M+1}{n+1}} \tag{10}$$

where the non-integral combinatoric is defined in terms of gamma functions.

We now propose another justification for a uniform prior from the point of view of an urn with two types of objects.

Suppose we have an urn with a known number of M red and white balls (or marked heads and tails) of which an unknown number R are red. The object is to infer R or equivalently R/M, since M is assumed known.

The urn may be sampled without replacement in a variety of ways. No matter what the sampling procedure, (except perhaps one that will exhaust the urn and here it is largely irrelevant) it seems illogical in this case to base a prior for R on the sampling procedure. Most rules for determining so-called reference priors when only a finite number of alternative are assumed in the presence of little or no knowledge usually assign each alternative equal prior probability. We shall do so here.

In the case of hypergeometric sampling we note that the chance of drawing t red balls out of the n sampled is

$$\Pr[T=t \mid n,B,R] = \frac{\binom{R}{t}\binom{M-R}{n-t}}{\binom{M}{n}}, \quad t=0,1,\ldots, \min(R,n) \qquad (11)$$

$$= 0, \quad \text{otherwise.}$$

For negative hypergeometric sampling i.e., sampling until t red balls are in hand and denoting N as the variable sample size, we obtain

$$\Pr[N=n \mid t,M,R] = \frac{\binom{n-1}{t-1}\binom{M-n}{R-t}}{\binom{M}{R}}, \quad \text{for } n=t,t+1,\ldots, \min(M,t+M-R). \qquad (12)$$

$$= 0, \quad \text{otherwise.}$$

Now, as already indicated, prior probabilities for a finite number of alternatives are usually assumed equal when little is known beforehand, so that

$$\Pr[R=r \mid M] = (M+1)^{-1}.$$

Consequently, since it is an easy matter to check that the likelihood of R of (12) is the same as in the previous case (11), then in either situation

$$\Pr[R=r \mid M,n,t] = \frac{\binom{r}{t}\binom{M-r}{n-t}}{\binom{M+1}{n+1}}, \quad r=t,\ldots, \min(M-n+t,n)$$

$$= 0, \quad \text{otherwise.}$$

We also note that for large M,

$$\Pr[T=t|n,M,r] \doteq \binom{n}{t}\left(\frac{R}{M}\right)^t \left(1-\frac{R}{M}\right)^{n-t} \tag{14}$$

$$\Pr[N=n|t,M,R] \doteq \binom{n-1}{t-1}\left(\frac{R}{M}\right)^t \left(1-\frac{R}{M}\right)^{n-t} \tag{15}$$

$$\Pr\left[\frac{R}{M} \leq z\right] \doteq \frac{\Gamma(n+2)}{\Gamma(t+1)\Gamma(n-t+1)} \int_0^z x^t(1-x)^{n-t}\,dx \quad . \tag{16}$$

As M grows and $RM^{-1} \to \theta$ the l.h.s. of (14), (15) and (16) tend to their respective r.h.s. Further the limiting posterior density of RM^{-1} is the one given for θ in (7) which emanates from the uniform prior and consequently differs from the posteriors recommended by most of the aforementioned Bayesians for θ, with the exception of Bayes and Jeffries. When M is large inferences on the fraction R/M derived from (16) will not differ appreciably from inferences using the posterior induced by the methods previously discussed unless n is very small.

3. Comments.

In much scientific, technical and medical experimentation where individual trials are binary, the parametric Bernoulli model assuming independent copies with its resulting likelihood function (1) is used as a basis for the analysis of the data. Although this is a useful and convenient paradigm for this type of trial it is not as appropriate in most instances as one that assumes only a finite number of trials can be made no matter how large. The first model assumes that there is some value θ which is the probability of "success" on each individual trial while the second entertains the notion that from a number of binary events, either having occurred or that potentially can occur, a certain number are observed which are in no way distinguishable from the rest before being observed. The latter model actually can include the first even when the first has some legitimate claim to describe the process--such as repeated tossing of the same coin. We assume the tosses are generated so that there is a sequence of heads and tails and then some fraction of the sequence is observed. One of the advantages for a Bayesian for the second approach is the greater simplicity (difficult though it is) in thinking about prior distributions for the observable quantities--say the number of heads out of the total rather than trying to focus on a distribution for values of θ.

The finite model that focuses on the fraction of successes in the finite population is basically a dependent Bernoulli model which leads to the hypergeometric likelihood, common to the particular cases of (11) and (12),

$$L(R) = \frac{R!(M-R)!}{(R-t)!(M-R-n+t)!} \quad . \tag{17}$$

Of course in many situations we are interested in the chance that the next observation is a success. This is acomplished by letting $M = t+1$ and is useful in determining what the chances are that a therapy already given to n ailing people more or less similar to you and having "cured" t of them will

also cure you of the ailment. A physician who wishes to know the chances that a particular fraction out of a given number say, M-n, that he is treating will be cured can calculate the chance that

$$\Pr\left[\frac{R-t}{M-n} \leq z\right] . \tag{18}$$

A pharmaceutical company or a government health organization may assume that the potential number of future cases is sufficiently large so that an asymptotic approximation is accurate enough to be adequately informed regarding the cure fraction of that finite, though not necessarily specified, number of these cases.

Recently, Stigler (1981) has argued that Bayes, himself, in his famous Scholium had actually presented a predictive argument for his uniform prior, which had been misinterpreted by critics. Although the argument attributed to Bayes by Stigler differs somewhat for that given here, it is tied together by the same predictive thread.

References

Akaike, H. (1978). A new look at the Bayes procedure. Biometrika, 65, 53-59.

Barnard, G.A., Jenkins, G.M. and Winsten, C.B. (1962). 'Likelihood inference and time series', (with discussion). J.R. Statist. Soc., A, 125, 321-372.

Box, G.E.P. and G.C. Tiao. (1973). Bayesian Inference in Statistical Analysis, Addison-Wesley Publishing Co.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). J. R. Statist Soc., B 41, 113-147.

Geisser, S. (1979). Discussion of Reference posterior distributions for Bayesian inference by J.M. Bernardo. J. R. Statist. Soc., B 41, 136-137.

Jeffreys, H. (1961). Theory of Probability. Oxford: University Press.

Stigler, S. (1982). Thomas Bayes's Bayesian inference. J.R. Statist. Soc. A, 145, 250-258.

Zellner, A. (1977). Maximal data information prior distributions. New Developments in the Application of Bayesian Methods, ed. A. Aykac and C. Brumat, North-Holland Publishing Co.

A Remark On A Model Criticism Technique

by

Seymour Geisser
University of Minnesota

Technical Report No. 409
October 1982

A Remark on a Model Criticism Technique

by
Seymour Geisser*
University of Minnesota

Box (1980) suggests a method for criticizing an entertained model consisting of data y, parameter set θ and assumptions A, which is structured via the relationship,

$$p(y, \theta|A) = p(y|\theta,A)\ p(\theta|A).$$

Here $p(y|\theta,A)$ is the joint probability function of the observations given θ and $p(\theta|A)$ is the prior probability function of θ. He notes that prior to the availability of the data one can compute

$$p(y|A) = \int p(y|\theta,A)\ p(\theta|A)\ d\theta$$

which he denotes as the predictive (marginal) density of the data. This, he claims, enables one to assess the credibility of the model for any observed set of data y_d by referring to $p(y_d|A)$ or to the density $p(g(y_d)|A)$ of some predictive checking function $g(y_d)$.

Basically he defines a test with significance level

$$\alpha = Pr\{p(y|A) < p(y_d|A)\}$$

to allow criticism of the model. He illustrates the concept by presenting several useful examples. In this note we shall present two examples which when taken in tandem throw some doubt on an uncritical use of this procedure.

Assume an i.i.d. sequence of Bernoulli trials with probability θ of success Suppose in this instance the prior denisty for θ is actually assumed to be uniform as in Bayes' original model. Then for a fixed number n of trials where y successes are observed, the predictive probability function of y

*Supported in part by NIH grant GM25271

is easily calculated to be

$$\Pr(y|A) = \frac{1}{n+1} \qquad y = 0, 1, \ldots, n. \tag{1}$$

i.e. uniform for all admissible values of y. Hence no significance test of the type advocated by Box,

$$\Pr\{p(y|A) < p(y_d|A)\} = \alpha \tag{2}$$

is available. Are we to conclude that Bayes' original model cannot be flawed? Or is it just that predictive model criticism fails here?

Suppose now we had used negative binomial sampling so that we terminated the experiment as soon as y successes were attained and consequently observed n trials. Here the predictive probability function of the number of trials is

$$\Pr(N=n|A) = \frac{y}{n(n+1)} \qquad n = y, y+1, \ldots . \tag{3}$$

The fact that the probability function is monotonically decreasing in n indicates that the Box procedure is workable i.e. if the observed $N = n_0$ is large enough relative to y, the model may be called into question. In fact,

$$\Pr[N \geq n_0] = \sum_{n=n_0}^{\infty} \frac{y}{n(n+1)} = \frac{y}{n_0} = \alpha, \tag{4}$$

where $\alpha = \hat{\theta}$, the MLE of θ. Are we then to conclude that predictive model criticism here succeeds only for small $\hat{\theta}$? Sampling until a fixed number of failures is attained results in criticism increasing with $\hat{\theta}$. In either case what aspect of the model is called into question other than Bayes' uniform prior? And, what is the meaning of calling this into question? Box has made an elegant suggestion for the problem of model criticism, but it should, like most statistical techniques, be used with caution and carefully interpreted within the context of its application.

REFERENCES

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. J.R. Statist. Soc. A, 143, 383-430.